Large Covariance Matrix Estimation under Heavy Tails via Regularized Huber's M-Estimation

Jialin Yu, *Student Member, IEEE*, Quan Wei, *Student Member, IEEE*, and Ziping Zhao, *Member, IEEE*

***Abstract*—XXX**

Keywords: Heavy-tailed distribution, High-dimensional statistics, Sparsity, M-estimation, Pairwise-difference approach

I. Introduction

[Covariance matrix estimation] A fundamental problem in Modern multivariate data analysis is the estimation of covariance matrices, which has widespread application across numerous fields, including statistics [2], biology [3]–[5], finance [6]–[8], signal processing [9]–[11], and machine learning [12]. The sample covariance matrix (SCM), which originates as the maximum likelihood estimator under a multivariate normal model, is simple, consistent and has a low asymptotic variance under various distributional assumptions. However, when the sample size is relatively small in comparison with the data dimension, the sample covariance can no longer benefit from its nice asymptotic properties.

[sparsity/high-dimensional covariance estimation methods, see Wei's introduction] When the data dimension is large, the covariance estimation problem becomes intractable. This is because the number of parameters to be estimated grows quadratically with the dimension of the covariance matrix. To reduce the number of parameters to be estimated, we need to introduce structural assumptions on the underlying covariance matrix, and one of the most popular assumptions is sparsity. A commonly used method for estimating sparse covariance matrices is called thresholding [13]–[15], which is to set small elements in the SCM to zeros. In particular, the soft thresholding covariance estimator is equivalent to the (unconstrained) ℓ_1-regularized sparse covariance estimator, which has been extensively studied (see [16]–[18]) and is proved to achieve the minimax optimal statistical rate under sub-Gaussian data. However, the ℓ_1 penalty introduces a non-negligible bias into the resulting estimator. To alleviate this bias effect, [1] substitutes the ℓ_1 penalty with a non-convex penalty, which results in a more refined estimator that achieves the oracle rate under sub-Gaussian data.

[why robustness/outliers] Despite the proceedings in developing high-dimensional covariance estimators, theoretical properties of large covariance estimators in the literature often hinge heavily on the Gaussian or sub-Gaussian assumption. However, such an assumption is very restrictive in some real-world scenarios, because the collected data is often contaminated by heavy-tailed noise. For instance, data from fields including biology [19] and finance [20] possess a heavy-tailed nature. The presence of heavy-tailed noise increases the frequency of outliers. An estimator that is robust against the outliers caused by heavy-tailed noise, evidenced by its better finite-sample performance than a non-robust estimator, is called a tail-robust estimator [21], [22]. The work of [23] inspires the design of tail-robust estimators in various statistics problems, including mean estimation [24], regression [25], and covariance estimation [26]. Those estimators are featured by tight non-asymptotic error bounds.

[Robust covariance: existing methods] In particular, tail-robust covariance estimators have been extensively studied. A line of work studies the robust-loss M-estimator [21], [27]–[29], with Huber's M-estimator [30] being the most representative. For instance, [29] uses Huber's M-estimators to estimate the first and second moments separately, which are then combined to obtain a robust covariance estimator. To avoid the accumulated error from combining the estimates of first and second moments, [21] proposes to use the pairwise difference approach [28] to directly estimate the covariance matrix. To optimize the performance of the robust-loss M-estimator, we need to carefully select the robustification parameter in its loss function. With a diverging robustification parameter adapted to the sample size, dimension and the noise level [31], Huber's M-estimators (and many other robust loss M-estimators) have achieved the optimal deviation bound in ℓ_∞ norm or spectral norm, assuming only a finite fourth moment for the distribution of data [21], [27], [28], [32]. A closely related covariance estimator uses truncation to eliminate outliers introduced by heavy-tailed noises [21], which is simpler and more computationally efficient than Huber's M-estimator, but still requires a selection of the robustification parameter. Another robust covariance estimator based on the median-of-means technique [33]–[35] avoids the robustification parameter in its formulation. This estimator randomly partitions the data into a prespecified number of groups, calculates the sample covariance matrices of those groups, and computes their median in each entry to obtain the final estimate. This method avoids the selection of robustification parameters and can be tuning-free, but requires a more restrictive assumption than a finite fourth moment, e.g. a finite sixth moment. There are also correlation-based methods that combine the estimation of correlation matrices with the robust estimation of marginal deviations to obtain a robust covariance estimator: By

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62001295 and in part by the Shanghai Sailing Program under Grant 20YF1430800. *(Corresponding author: Ziping Zhao.)* This paper was presented in part at the 48th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Rhodes Island, Greece, June 4–10, 2023 [1].

exploiting a bijective mapping between the correlation matrix and Kendall's tau or Spearman's rho dependence measures, [36] and [37] both proposed rank-based correlation estimators for heavy-tailed elliptical distributions, which can be combined with marginal standard deviations estimated via various robust methods [22], [29], [38] to obtain a robust covariance estimator. The aforementioned estimators are mostly constructed and analyzed in an entrywise manner. There are also robust covariance estimators that are studied from a spectrumwise perspective [26], [39].

[why robust + sparsity/high-dimensional] To develop robust covariance estimators for high-dimensional covariance matrices, we need to introduce structural assumptions. We will first focus on the sparsity assumption mentioned earlier. A common robust sparse covariance estimation method is to first introduce a robust pilot estimator[1] as a robust substitution of the sample covariance matrix, then apply thresholding to the pilot estimator as in [29] and [40], or apply ℓ_1 regularization to the pilot estimator as in [41]. In particular, for the robust pilot estimator, [29] considers the median-of-means estimator, Huber's M-estimator and the rank-based estimator; [41] uses their proposed quantile-based covariance estimator as the robust pilot estimator; and [40] uses Maronna's estimator as the robust pilot estimator. Apart from the sparsity in the covariance matrix itself, we can also assume the sparsity in its spectral, which is usually called the low rank assumption. To utilize the latent low rank structure and obtain robust estimators with a tight deviation under the Frobenius norm, [27] considers a trace-norm-regularized M-estimator proposed by [42], but substitutes the sample covariance it use with a robust pilot estimator.

As illustrated above, all existing robust sparse covariance estimation methods follow a two-step procedure, where a robust pilot estimator is introduced in the first step, and thresholding or ℓ_1 regularization is applied to the pilot estimator [29], [40], [41] in the second step. To obtain a robust pilot estimator, [29] uses Huber's M-estimators to estimate the first and second moments separately, which are then combined to obtain a robust covariance estimator. To avoid the accumulated error from combining the estimates of first and second moments, [21] uses the pairwise difference approach to directly obtain a covariance estimator. However, the estimators proposed by [29] and [21] are not guaranteed to be positive-definite, which is important for the covariance estimation problem. To simultaneously achieve sparsity and positive-definiteness, [41] adds a positive-definite constraint to the ℓ_1 regularization step. In general, all existing tail-robust sparse covariance estimation procedures proceed in two separate steps, with the first step being a "robustification" procedure which results in a robust pilot estimator, and the second step being a "sparsification" procedure based on this robust pilot estimator, which results in a (positive-definite) robust sparse covariance estimator. In this two-step procedure, robustness and sparsity are considered separately, which results in an accumulated statistical error. However, a direct combination of the two steps remains to be explored, which, as we anticipated, turns out to be a one-step estimator that jointly considers robustness and sparsity, enjoys an optimal performance theoretically, and excels among other methods numerically.

[1]Here we refer to the pilot estimator as a tail-robust estimator that achieves a certain deviation bound in ℓ_∞-norm or spectral norm, which follows the terminology in [29].

A. Contributions

In this paper, we investigate the robust sparse covariance estimation problem in the high-dimensional regime. In existing literature, all tail-robust sparse covariance estimation procedures proceed in two separate steps bridged by an intermediate pilot estimator. In this paper, we will combine the two separate steps into a single-step Huber-loss ℓ_1-penalized sparse covariance estimator. We will refer to it as the regularized Huber's-M estimator. The main contributions of this paper are summarized as follows:

- We propose the regularized Huber's-M estimator for robust sparse covariance estimation problem. By using pairwise-difference approach in the loss function, the proposed covariance estimation method not only avoids a separate estimation of the population mean, but also exploits the sample data more efficiently, which is especially beneficial in the high-dimensional regime.
- We clearly demonstrate the statistical properties for the proposed estimator: Assuming only a finite fourth moment for the data distribution, our estimator achieves the minimax optimal rate under both ℓ_1 norm and the Frobenius norm. In comparison, the multi-stage sparse covariance estimator proposed by [41] overcomes heavy-tailed high-dimensional data and achieves the minimax optimal statistical rate as our's does, but their result hinges on an additional elliptical-shape assumption that is only known to hold for pair-elliptically distributed data.
- The performance of our proposed estimator is compared with other methods, which validates the superiority of our estimator over existing methods and supports the theory on its statistical rate.

B. Organization

The rest of the paper is organized as follows. In Section III, we introduce the pairwise-difference approach and explain our formulation for the robust sparse covariance matrix estimation problem. In Section V and VI, we present theoretical results, including the statistical convergence rates of the proposed estimator under ℓ_1 norm and Frobenius norm. In Section VII, we will evident via simulation that the proposed estimator excels among other methods. We conclude by discussions in Section VIII. The proofs of all theoretical results are given in the Appendix.

C. Notation

The following notation is adopted. Standard lower-case or upper-case letters stand for scalars and boldface lower-case (upper-case) letters denote vectors (matrices). Both X_{ij} and

$[\mathbf{X}]_{ij}$ denote the (i,j)-th entry of the matrix $\mathbf{X}$. $\mathbb{R}_+$ denotes the set of non-negative real numbers, $\mathbb{R}^{m\times n}$ denotes the set of real $m\times n$ matrices. $\mathbf{0}$ and $\mathbf{1}$ stand for the all-zero and all-one vector/matrix, respectively. $\mathbf{I}$ stands for the identity matrix. $\mathbf{X}\succ\mathbf{0}$ ($\mathbf{X}\succeq\mathbf{0}$) means $\mathbf{X}$ is positive definite (semidefinite). $\mathbf{x}\geq\mathbf{0}$ denotes each element of $\mathbf{x}$ is non-negative.

Let $\|\mathbf{X}\|_F=\sqrt{\sum_{ij}X_{ij}^2}$ and $\|\mathbf{X}\|_1=\sum_{ij}|X_{ij}|$ denote Frobenius norm and the ℓ_1 norm, respectively. Let $\|\mathbf{X}\|_\infty=\max_{k,l}|X_{kl}|$ denote the ℓ_∞ norm and $\|\mathbf{X}\|_{1,\text{off}}=\sum_{k\neq l}|X_{kl}|$ denote the sum-absolute-value norm for all entries and for off-diagonals. We write $[d]$ for the set $\{1,2,\ldots,d\}$ and $\lfloor x\rfloor$ for the largest integer not exceeding x. For an index set $\mathcal{S}$, we use $|\mathcal{S}|$ to denote its cardinality, $\overline{\mathcal{S}}$ to denote its complement. Use $\mathbf{X}_{\mathcal{S}}$ to denote the matrix whose (i,j)-th entry is equal to X_{ij} if $(i,j)\in\mathcal{S}$, and zero otherwise. Let $\partial f(\cdot)$ denote the subgradient of a function f.

Let $\text{sign}(x)$ denote the sign of variable x, i.e., $\text{sign}(x)=x/|x|$ for $x\neq 0$ and $\text{sign}(0)=0$. For functions $f(n)$ and $g(n)$, we denote $f(n)\lesssim g(n)$ or $f(n)=O(g(n))$ if $f(n)\leq Cg(n)$, $f(n)\gtrsim g(n)$ or $f(n)=\Omega(g(n))$ if $f(n)\geq cg(n)$, and denote $f(n)\asymp g(n)$ if $cg(n)\leq f(n)\leq Cg(n)$ for some positive absolute constants c and C.

Let $\mathbb{E}[x]$ and $\text{Var}(x)$ denote the expectation and variance of a random variable x, respectively. Let $\Pr(\mathcal{E})$ denote the probability of a random event $\mathcal{E}$.

II. Prior Art: Non-robust Covariance Estimation

A. Covariance fitting loss

Let $\mathbf{x}_i$, $i=1,\ldots,n$ be samples of a d-dimensional random variable $\boldsymbol{x}$. The sample covariance matrix $\mathbf{S}$ is computed as

$$\mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T$$

with $\mathbf{m}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ being the sample mean. In the literature, a common approach for covariance matrix estimation is based on the regularized least squares method [1], [16]–[18]

$$l_2(\mathbf{\Sigma}):=\frac{1}{2}\|\mathbf{S}-\mathbf{\Sigma}\|_F^2+\lambda\|\mathbf{\Sigma}\|_{1,\text{off}}. \tag{1}$$

where sample covariance matrix $\mathbf{S}$ serves as a pilot estimator. However, this formulation is non-robust [21]. The reason behind can be explained in the following subsection.

B. A mean vector estimation interpretation

Suppose

$$\mathbf{x}_i=\boldsymbol{\mu}+\mathbf{e}_i$$

where $\boldsymbol{\mu}$ is the unknown mean.

Let $N:=n(n-1)/2$ and define the paired data

$$\{\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_N\}=\{\boldsymbol{x}_i-\boldsymbol{x}_j\}_{1\leq i<j\leq n},$$

and consider

$$\begin{aligned}\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T&=\frac{1}{2}(\mathbf{e}_i-\mathbf{e}_j)(\mathbf{e}_i-\mathbf{e}_j)^T\\&=\frac{1}{2}\left(\mathbf{e}_i\mathbf{e}_i^T+\mathbf{e}_j\mathbf{e}_j^T\right)-\frac{1}{2}\left(\mathbf{e}_j\mathbf{e}_i^T+\mathbf{e}_i\mathbf{e}_j^T\right)\\&=\mathbf{\Sigma}^*+\mathbf{E}_i\end{aligned}$$

where $\mathbf{E}_i=\frac{1}{2}\left(\mathbf{e}_i\mathbf{e}_i^T+\mathbf{e}_j\mathbf{e}_j^T\right)-\mathbf{\Sigma}^*-\frac{1}{2}\left(\mathbf{e}_j\mathbf{e}_i^T+\mathbf{e}_i\mathbf{e}_j^T\right)$. The pairwise difference approach can get rid off the estimation of the mean $\boldsymbol{\mu}$.

When $\mathbf{x}$ is assumed to follow a sub-Gaussian distribution, the elements of $\mathbf{E}_i$ follow a light-tailed distribution. This is because the product of subgaussian random variables $\mathbf{e}_i$ and $\mathbf{e}_j$ (i.e. $\mathbf{e}_i\mathbf{e}_i^T$ or $\mathbf{e}_i\mathbf{e}_j^T$) follow a light-tailed distribution (see Lemma 2.7.7 in [43]), and as a summation of such products, $\mathbf{E}_i$ follows a light-tailed distribution as well. Therefore, based on a mean vector estimation perspective and applying a squared loss function for elements of $\mathbf{E}_i$, we obtain the following loss function:

$$\frac{1}{2N}\sum_{i=1}^{N}\sum_{k,l=1}^{d}(\Sigma_{kl}-\frac{1}{2}y_{ik}y_{il})^2$$

Then, we have

$$\begin{aligned}&\frac{1}{2N}\sum_{i=1}^{N}\sum_{k,l=1}^{d}(\Sigma_{kl}-\frac{1}{2}y_{ik}y_{il})^2\\=&\frac{1}{2N}\sum_{i=1}^{N}\left\|\mathbf{\Sigma}-\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T\right\|_F^2\\=&\frac{1}{2}\left(\langle\mathbf{\Sigma},\mathbf{\Sigma}\rangle-2\left\langle\mathbf{\Sigma},\frac{1}{2N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^T\right\rangle\right.\\&\left.+\frac{1}{N}\sum_{i=1}^{N}\left\langle\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T,\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T\right\rangle\right)\\=&\frac{1}{2}\left(\langle\mathbf{\Sigma},\mathbf{\Sigma}\rangle-2\langle\mathbf{\Sigma},\mathbf{S}\rangle+\frac{1}{N}\sum_{i=1}^{N}\left\langle\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T,\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T\right\rangle\right)\\=&\frac{1}{2}\|\mathbf{\Sigma}-\mathbf{S}\|_F^2+\underbrace{\frac{1}{2}\left(-\langle\mathbf{S},\mathbf{S}\rangle+\frac{1}{N}\sum_{i=1}^{N}\left\langle\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T,\frac{1}{2}\boldsymbol{y}_i\boldsymbol{y}_i^T\right\rangle\right)}_{=const.}\end{aligned}$$

where in the third line we have used the relation $\mathbf{S}=\frac{1}{2N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^T$ (see the proof in the Appendix).

The above mathematical derivation reveals the classically applied least squares method is based on the squared loss for the error term $\mathbf{E}_i$. It is generally acknowledged that the regularized least squares method is efficient only when the error follows a light-tailed distribution [32].

III. Problem Formulation

When $\boldsymbol{e}_i$ is heavy-tailed with a finite fourth moment, $\boldsymbol{E}_i$ will also be heavy-tailed in general, and with a finite second moment. Following the perspective of [23] and applying a Huber loss function for elements of $\mathbf{E}_i$, we propose to use the following robust loss function:

$$L_\alpha(\mathbf{\Sigma}):=\sum_{k,\ell}\frac{1}{N}\sum_{i=1}^{N}\rho_\alpha(\Sigma_{k\ell}-y_{ik}y_{i\ell}/2), \tag{2}$$

with $\rho_\alpha:\mathbb{R}\to\mathbb{R}_+$ a Huber loss function defined as

$$\rho_\alpha(x)=\begin{cases}x^2/2 & \text{if } |x|\leq\alpha\\ \alpha|x|-\alpha^2/2 & \text{if } |x|>\alpha\end{cases}$$

for some non-negative robustification parameter α. Based on $L_\alpha(\mathbf{\Sigma})$, [21] proposed a Huber's M-estimator for covariance:

$$\widehat{\mathbf{\Sigma}}^H = \arg\min_{\mathbf{\Sigma}\in\mathbb{R}} L_\alpha(\mathbf{\Sigma}). \tag{3}$$

When $\alpha \to \infty$, the Huber loss in (2) becomes the squared loss, and

$$\widehat{\mathbf{\Sigma}}^H = \arg\min_{\mathbf{\Sigma}\in\mathbb{R}} \sum_{k,\ell} \frac{1}{2N} \sum_{i=1}^{N} (\Sigma_{k\ell} - y_{ik}y_{i\ell}/2)^2$$

which has a closed-form solution that matches the sample covariance

$$\begin{aligned}\widehat{\mathbf{\Sigma}}^H &= \frac{1}{2N}\sum_{i=1}^{N} \boldsymbol{y}_i\boldsymbol{y}_i^T \\ &= \frac{1}{2N}\sum_{1\le i<j\le n} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T = \boldsymbol{S}.\end{aligned}$$

The sample covariance is not robust. To trade off bias against robustness, we need to carefully select the robustification parameter α [31]. Taking $\alpha \asymp \sqrt{n/\log d}$, it can be shown that $\left\|\widehat{\mathbf{\Sigma}}^H - \mathbf{\Sigma}^*\right\|_\infty \lesssim \sqrt{\log d/n}$ with high probability, which is optimal under ℓ_∞ norm [21]. Nevertheless, $\widehat{\mathbf{\Sigma}}^H$ is generally not a sparse estimation when $\mathbf{\Sigma}^*$ is sparse.

A. The proposed estimator

Rather than applying thresholding to $\widehat{\mathbf{\Sigma}}^H$, we propose to induce sparsity by directly adding an ℓ_1 penalty to (3), that is,

$$\widehat{\mathbf{\Sigma}} \in \arg\min_{\mathbf{\Sigma}} \left\{ L_\alpha(\mathbf{\Sigma}) + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \right\}. \tag{4}$$

To further obtain positive-definiteness, we propose to add a log-determinant barrier function to (4), which results in the following estimator:

$$\widehat{\mathbf{\Sigma}}^+ \in \arg\min_{\mathbf{\Sigma}} \left\{ L_\alpha(\mathbf{\Sigma}) - \tau \log\det \mathbf{\Sigma} + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \right\}, \tag{5}$$

where $\tau > 0$ is the barrier parameter and $\log\det\mathbf{\Sigma}$ is defined to be negative infinity for $\mathbf{\Sigma}$ not positive definite. The log-determinant barrier term ensures the existence of a positive definite solution [16]. We will demonstrate that $\widehat{\mathbf{\Sigma}}^+$ retains the desirable properties of $\widehat{\mathbf{\Sigma}}$, including robustness, sparsity, and minimax optimality in the statistical rate.

Remark 1. In the literature, another popular robust estimation method is to replace the pilot estimator in (1) by a robust covariance estimator. (1) However, compared with our method, this is a two-step method. It can be computationally expensive and engenders an accumulated estimation error [21]. ~~(2) they need to estimate the mean but we get rid of this mean estimation step.~~

IV. Computational Algorithm

A. A convex ADMM algorithm

Recall that $\log\det\mathbf{\Sigma}$ is defined to be negative infinity for $\mathbf{\Sigma}$ not positive definite. We want to optimize the following problem:

$$\underset{\mathbf{\Sigma}}{\text{minimize}} \left\{ \frac{1}{N}\sum_{j=1}^{N} \rho_\alpha(\mathbf{\Sigma} - \boldsymbol{R}_j) - \tau\log\det\mathbf{\Sigma} + \lambda\|\mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{6}$$

where $\rho_\alpha(\mathbf{\Sigma} - \boldsymbol{R}_j) = \sum_{k,\ell}\rho_\alpha(\Sigma_{k\ell} - R_{jkl})$. Given that the Huber loss is not strongly convex, simply applying a projected sub-gradient algorithm to problem (6) results in an $O(1/\sqrt{K})$ convergence rate after K iterations, which is computationally inefficient.

To alleviate the difficulty caused by the Huber loss function, a feasible way is to introduce a Majorization-Minimization (MM) framework, which replaces ρ_α with a quadratic upper bound in each iteration. To be more specific, given the ith update $\mathbf{\Sigma}^i$, the quadratic upper bound can be defined as

$$h_j\left(\mathbf{\Sigma}\left|\mathbf{\Sigma}^i\right.\right) = \left\|\mathbf{\Sigma} - \mathbf{\Sigma}^i + \nabla\rho_\alpha\left(\mathbf{\Sigma}^i - \boldsymbol{R}_j\right)\right\|_{\text{F}}^2/2$$

It can be seen that $\nabla\rho_\alpha\left(\mathbf{\Sigma}^i - \boldsymbol{R}_j\right) = \nabla h_j\left(\mathbf{\Sigma}^i\left|\mathbf{\Sigma}^i\right.\right)$, and $\rho_\alpha(\mathbf{\Sigma} - \boldsymbol{R}_j) \le h_j\left(\mathbf{\Sigma}\left|\mathbf{\Sigma}^i\right.\right)$ for all $\sigma \in \mathbb{R}$. Applying this upper bound to each and every term in (6), we obtain

$$\mathbf{\Sigma}^{i+1} = \min_{\mathbf{\Sigma}} \left\{ \frac{1}{2}\left\|\mathbf{\Sigma} - \widetilde{\boldsymbol{R}}\right\|_{\text{F}}^2 - \tau\log\det\mathbf{\Sigma} + \lambda\|\mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{7}$$

where $\widetilde{\boldsymbol{R}} = \mathbf{\Sigma}^i - \frac{1}{N}\sum_{j=1}^{N}\left[\nabla\rho_\alpha\left(\mathbf{\Sigma}^i - \boldsymbol{R}_j\right)\right]$. In each iteration, problem (7) can be solved using an algorithm proposed by Rothman [16].

Before the first iteration, we initialize $\mathbf{\Sigma}^0 = \boldsymbol{I}$ for simplicity. The MM algorithm for solving (6) is summarized in Algorithm 1.

Algorithm 1: The MM algorithm for solving (8).

Input: $\{\mathbf{R}_j\}_{j=1}^N$, τ, λ;
Initialize $\mathbf{\Sigma}^0 = \boldsymbol{I}$, $i = 0$;
repeat
  update $\widetilde{\boldsymbol{R}} = \mathbf{\Sigma}^i - \frac{1}{N}\sum_{j=1}^{N}\left[\nabla\rho_\alpha\left(\mathbf{\Sigma}^i - \boldsymbol{R}_j\right)\right]$;
  obtain $\mathbf{\Sigma}^{i+1}$ by solving 7 with Rothman's algorithm;
  $i = i + 1$;
until *convergence*;
Output: $\mathbf{\Sigma}^i$.

However, Rothman's algorithm does not have a theoretical guarantee for its convergence rate.

We propose to use the alternating direction method of multipliers (ADMM) to solve this problem. We first introduce a new variable $\mathbf{\Omega}$ and an equality constraint as follows:

$$\underset{\mathbf{\Sigma},\mathbf{\Omega},\mathbf{\Sigma}=\mathbf{\Omega}}{\text{minimize}} \left\{ \frac{1}{N}\sum_{j=1}^{N} \rho_\alpha(\mathbf{\Sigma} - \boldsymbol{R}_j) - \tau\log\det\mathbf{\Omega} + \lambda\|\mathbf{\Sigma}\|_{1,\text{off}} \right\} \tag{8}$$

The solution to (8) gives the solution to (6). To deal with the constraint $\boldsymbol{\Sigma} = \boldsymbol{\Omega}$ in (8), consider the augmented Lagrangian function for some given penalty parameter ρ:

$$\begin{aligned} & L\left(\boldsymbol{\Sigma}, \boldsymbol{\Omega}, \boldsymbol{Y}\right) \\ & = \frac{1}{N} \sum_{j=1}^{N} \rho_{\alpha}(\boldsymbol{\Sigma} - \boldsymbol{R}_j) - \tau \log\det \boldsymbol{\Omega} + \lambda \left\|\boldsymbol{\Sigma}\right\|_{1,\text{off}} \\ & \qquad + \left\langle \boldsymbol{Y}, \boldsymbol{\Sigma} - \boldsymbol{\Omega} \right\rangle + \frac{\rho}{2} \left\|\boldsymbol{\Sigma} - \boldsymbol{\Omega}\right\|_{\text{F}}^2, \end{aligned}$$

where $\boldsymbol{Y}$ is the Lagrange multiplier. In each iteration, we sequentially update $\boldsymbol{\Sigma}^{i+1}$, $\boldsymbol{\Omega}^{i+1}$ and $\boldsymbol{Y}^{i+1}$ as follows

$$\begin{aligned} \boldsymbol{\Sigma}^{i+1} &= \arg\min_{\boldsymbol{\Sigma}} L\left(\boldsymbol{\Sigma}, \boldsymbol{\Omega}^i, \boldsymbol{Y}^i\right) \\ &= \arg\min_{\boldsymbol{\Sigma}} \frac{1}{N} \sum_{j=1}^{N} \rho_{\alpha}(\boldsymbol{\Sigma} - \boldsymbol{R}_j) + \lambda \left\|\boldsymbol{\Sigma}\right\|_{1,\text{off}} \\ &\qquad + \left\langle \boldsymbol{Y}^i, \boldsymbol{\Sigma} - \boldsymbol{\Omega}^i \right\rangle + \frac{\rho}{2} \left\|\boldsymbol{\Sigma} - \boldsymbol{\Omega}^i\right\|_{\text{F}}^2 \\ \boldsymbol{\Omega}^{i+1} &= \arg\min_{\boldsymbol{\Omega}} L\left(\boldsymbol{\Sigma}^{i+1}, \boldsymbol{\Omega}, \boldsymbol{Y}^i\right) \\ &= \arg\min_{\boldsymbol{\Omega}} \frac{\rho}{2} \left\|\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Omega}\right\|_{\text{F}}^2 + \left\langle \boldsymbol{Y}^i, \boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Omega} \right\rangle \\ &\qquad - \tau \log\det \boldsymbol{\Omega} \\ &= \arg\min_{\boldsymbol{\Omega} \succ \boldsymbol{0}} \frac{\rho}{2} \left\| \boldsymbol{\Sigma}^{i+1} + \frac{1}{\rho}\boldsymbol{Y}^i - \boldsymbol{\Omega} \right\|_{\text{F}}^2 - \tau \log\det \boldsymbol{\Omega} \\ \boldsymbol{Y}^{i+1} &= \boldsymbol{Y}^i + \frac{1}{\rho}\left(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Omega}^{i+1}\right). \end{aligned}$$

When updating $\boldsymbol{\Sigma}^{i+1}$, the objective function is separable, and for each (k, l), the entry Σ_{kl}^{i+1}can be obtained separately as a solution to the following equation:

$$\begin{aligned} & \frac{1}{N} \sum_{j=1}^{N} \rho'_{\alpha}(\Sigma_{kl}^{i+1} - R_{jkl}) + I(k \neq l) \cdot \lambda\xi \\ & \qquad + Y_{kl}^i + \rho\Sigma_{kl}^{i+1} - \rho\Omega_{kl}^i = 0, \quad \xi \in \partial \left|\Sigma_{kl}^{i+1}\right|, \end{aligned} \tag{9}$$

which can be easily computed using the bisection method that converges linearly. Also note that the above update can be parallelized when necessary.

To update $\boldsymbol{\Omega}^{i+1}$, consider the spectral decomposition of $\boldsymbol{\Sigma}^{i+1} + \frac{1}{\rho}\boldsymbol{Y}^i$. That is, $\boldsymbol{\Sigma}^{i+1} + \frac{1}{\rho}\boldsymbol{Y}^i = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^T$, where

$$\boldsymbol{E} = \begin{bmatrix} e_1 & & \\ & \ddots & \\ & & e_d \end{bmatrix},$$

then the update for $\boldsymbol{\Omega}^{i+1}$ can be written as

$$\boldsymbol{\Omega}^{i+1} = \boldsymbol{U}\boldsymbol{E}'\boldsymbol{U}^T \tag{10}$$

where

$$\boldsymbol{E}' = \begin{bmatrix} \phi\left(e_1\right) & & \\ & \ddots & \\ & & \phi\left(e_d\right) \end{bmatrix},$$

with $\phi(x) \coloneqq \frac{\sqrt{x^2 + \frac{4\tau}{\rho}} + x}{2}$.

Before the first iteration, we initialize $\boldsymbol{\Omega}^0 = \boldsymbol{I}$ and $\boldsymbol{Y}^0 = \boldsymbol{0}$ for simplicity. The ADMM algorithm for solving (8) is summarized in Algorithm 2.

Algorithm 2: The ADMM algorithm for solving (8).

Input: $\{\mathbf{R}_j\}_{j=1}^N$, τ, λ, ρ;
Initialize $\boldsymbol{\Omega}^0 = \boldsymbol{I}$, $\boldsymbol{Y}^0 = \boldsymbol{0}, i = 0$;
repeat
  obtain $\boldsymbol{\Sigma}^{i+1}$ by solving 9;
  obtain $\boldsymbol{\Omega}^{i+1}$ using (10);
  update $\boldsymbol{Y}^{i+1} = \boldsymbol{Y}^i + \frac{1}{\rho}\left(\boldsymbol{\Sigma}^{i+1} - \boldsymbol{\Omega}^{i+1}\right)$;
  $i = i + 1$;
until *convergence*;
Output: $\boldsymbol{\Omega}^i$.

As illustrated in [44], with K iterations, this algorithm enjoys an $O(1/K)$ convergence rate, which is significantly faster than the projected sub-gradient algorithm.

B. A convex BMM algorithm

V. Main Results

In this section, we first introduce some necessary technical assumptions for the theoretical analysis, then establish the statistical convergence rates of the proposed covariance estimator under both ℓ_1 and Frobenius norm.

A. Assumptions

We denote the true covariance matrix by $\mathbf{\Sigma}^*$. Let $\mathcal{S} = \{(i,j) \mid \Sigma^*_{ij} \neq 0\}$ be the support set of $\mathbf{\Sigma}^*$ and s be its cardinality, i.e., $s = |\mathcal{S}|$. In the following, we impose some mild conditions on the distribution of the i.i.d. observations $\mathbf{x}_i$, $i = 1, \ldots, n$ and the structure of true covariance matrix $\mathbf{\Sigma}^*$.

Assumption 2 (fourth moment assumption)**.** *$\mathbf{x}_i \in \mathbb{R}^d$ is a heavy-tailed random vector such that $\mathbb{E}\left[|x_{ik} - \mathbb{E}[x_{ik}]|^4\right] \leq \sigma^2$ for all $1 \leq k \leq d$ with some positive constant σ, which implies that there exists $K > 0$ depending only on σ, such that Var$((x_{ik} - x_{jk})(x_{il} - x_{jl})/2) \leq K$ for all $k, l \in [d]$ and $i \neq j$.*

Remark 3. $\mathbb{E}\left[|x_{ik} - \mathbb{E}[x_{ik}]|^4\right] \leq \sigma^2$ implies that the fourth moment of the paired data $\{\boldsymbol{x}_i - \boldsymbol{x}_j\}_{1 \leq i < j \leq n}$ is also bounded, i.e. $\mathbb{E}\left[(x_{ik} - x_{jk})^4\right] \leq 4K$ with a constant $K > 0$ depending on σ. Hence for all $k, l \in [d]$, we have

$$\begin{aligned}&\text{Var}\left((x_{ik} - x_{jk})(x_{il} - x_{jl})/2\right)\\ &= -\left(\Sigma^*_{kl}\right)^2 + \mathbb{E}\left[(x_{ik} - x_{jk})^2 (x_{il} - x_{jl})^2\right]/4 \leq K.\end{aligned}$$

Also note that a scaling scheme of K with respect to d is implicitly assumed. In other words, K might also depend on d.

The fourth moment assumption in Assumption 2 is typical for the study of tail-robust covariance estimation, which is adopted in [4], [15], [21], [29].

Assumption 4 (positive-definiteness)**.** *The true covariance matrix satisfies $\mathbf{\Sigma}^* \succ \mathbf{0}$.*

The assumption on positive-definiteness is standard in existing literature on large covariance estimation problems. [16]–[18].

To analyze the statistical properties of the proposed estimator, we first define a "good" event regarding the local strong convexity of the empirical Huber loss over a local ℓ_∞ ball:

Definition 5. Given $\mathbb{B}^\infty(r) := \left\{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_\infty \leq r\right\}$, define $\mathcal{E}_1(r, \kappa)$ as the following event: For all $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(r)$,

$$\langle \nabla L_\alpha(\mathbf{\Sigma}_1) - \nabla L_\alpha(\mathbf{\Sigma}_2), \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \rangle \geq \kappa \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_{\mathrm{F}}^2.$$

In the next subsection, we will show that the Huber-based loss function $L_\alpha(\mathbf{\Sigma})$ is locally strongly convex i.e. the above event holds in an ℓ_∞ neighborhood of $\mathbf{\Sigma}^*$. By comparison, the squared loss is globally strongly convex, which makes it a lot easier to analysis than the Huber-based loss function.

B. Statistical Theory

Our theoretical results on the convergence rates will be provided in this section. We first provide a deterministic version of our convergence theorem in the following Proposition.

Proposition 6. *Assume $\alpha > 6\lambda s^{1/2}$ and condition on the event $\mathcal{E}_1(\alpha/2, 1/2) \cap \left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2\right\}$, we have*

$$\left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_F \leq 3\lambda s^{1/2} \quad \textit{and} \quad \left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_I \leq 12\lambda s.$$

Proposition 6 gives the deterministic version of our convergence theorem. In particular, it demonstrates rate under Frobenius norm and ℓ_1 norm. In the following discussions we will analyze the probabilities of the condition $\mathcal{E}_1(\alpha/2, 1/2)$

and the condition $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2$, respectively.

Strong convexity is important for the study of M-estimators. However, in our formulation (5), the Huber loss is not globally strongly convex. Therefore, we need to characterize the local strong convexity within an ℓ_∞ neighborhood of $\mathbf{\Sigma}^*$ with the following Proposition.

Proposition 7 (local strong convexity)**.** *Suppose that Assumption 2 holds. Assume $\alpha \asymp \sqrt{Kn/\log d}$. Then, for any $\kappa \in (0,1)$, with $n \gtrsim \log d$,*

$$\langle \nabla L_\alpha(\mathbf{\Sigma}_1) - \nabla L_\alpha(\mathbf{\Sigma}_2), \mathbf{\Sigma}_1 - \mathbf{\Sigma}_2 \rangle \geq \kappa \|\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2\|_F^2$$

holds uniformly for all $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(\alpha/2)$ with at least $1 - 2/d$ probability.

Proposition 7 implies that event $\mathcal{E}_1(\alpha/2, \kappa)$ happens with high probability. In the following Lemma, we will analyze the probability of the second condition $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2$.

Lemma 8. *Suppose that Assumption 2 hold. Then,*

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \leq \sqrt{12K \log d/n} + 4\alpha \log d/n + K/\alpha \quad (11)$$

holds with at least $1 - 2/d$ probability.

Further, let $\alpha \asymp \sqrt{Kn/\log d}$, $\tau \lesssim \sqrt{K \log d/n} \cdot \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty^{-1}$ and take $\lambda \asymp \sqrt{K \log d/n}$. Then,

$$\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2$$

holds with at least $1 - 2/d$ probability.

Now we present our statistical theory for the proposed estimator. We will study the error bounds under Frobenius norm and ℓ_1 norm, which match the minimax optimal rates for sparse covariance estimation [45], [46].

Theorem 9 (minimax-optimal rates)**.** *Suppose that Assumptions 2 and 4 hold. Let $\alpha \asymp \sqrt{Kn/\log d}$, $\tau \lesssim \sqrt{K \log d/n} \cdot \left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty^{-1}$ and take $\lambda \asymp \sqrt{K \log d/n}$. If the sample size satisfies $n \gtrsim s^{1/2} \log d$, then*

$$\left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_F \lesssim \sqrt{\frac{Ks \log d}{n}}$$
$$\text{and} \quad \left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K \log d}{n}}$$

hold with at least $1 - 4/d$ probability.

In comparison, under the Frobenius norm, the multi-stage sparse covariance estimator proposed by [41] achieves the minimax optimal statistical rate as our's does, but in addition to the fourth moment assumption (i.e. Assumption 2), their result hinges on an elliptical-shape condition that is only known to hold for pair-elliptically distributed data. In the rest of the literature, the Frobenius norm bound is barely discussed, even in some papers that have proposed important methodologies for robust sparse covariance estimation [29]. We are also the first to include deviation analysis for robust sparse covariance estimators under ℓ_1 norm, and our deviation bound matches the minimax rate in [45].

VI. A Constrained Robust Covariance Estimator

In existing literature, there are two ways of achieving positive-definiteness. One is to add a log-determinant barrier term, which is studied in the previous section. The other one is to add an eigenvalue constraint as illustrated in [18]. Similarly, for robust covariance estimation, consider adding the eigenvalue constraint to 4, and we obtain the following estimator:

$$\widehat{\mathbf{\Sigma}}_\epsilon \in \arg\min_{\mathbf{\Sigma} \succeq \epsilon \boldsymbol{I}} \left\{ L_\alpha(\mathbf{\Sigma}) + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \right\}, \quad (12)$$

where $\epsilon \geq 0$ lower bounds the minimum singular value of $\widehat{\mathbf{\Sigma}}_\epsilon$. Now we present our statistical theory for $\widehat{\mathbf{\Sigma}}_\epsilon$.

Theorem 10. *Suppose that Assumptions 2 and 4 hold. Let $\alpha \asymp \sqrt{Kn/\log d}$ and take $\lambda \asymp \sqrt{K \log d/n}$. If ϵ is smaller than the minimum singular value of $\mathbf{\Sigma}^*$ and the sample size satisfies $n \gtrsim \sqrt{s} \log d$, then there is a unique minimizer for problem (12). Let $\widehat{\mathbf{\Sigma}}_\epsilon$ denote this minimizer, then*

$$\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_F \lesssim \sqrt{\frac{Ks \log d}{n}}$$
$$\text{and} \quad \left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K \log d}{n}}$$

hold simultaneously with at least $1 - 4/d$ probability.

The constrained robust covariance estimator $\widehat{\mathbf{\Sigma}}_\epsilon$ is inherently related to our proposed $\widehat{\mathbf{\Sigma}}^+$: Roughly speaking, $\widehat{\mathbf{\Sigma}}_\epsilon$ can be viewed as the limit of $\widehat{\mathbf{\Sigma}}^+$ for parameter $\tau \to 0$ in (5). Although $\widehat{\mathbf{\Sigma}}_\epsilon$ is constructed to guarantee positive-definiteness from a different angle, it can be seen from Theorem 10 that the statistical properties for $\widehat{\mathbf{\Sigma}}_\epsilon$ are actually the same as $\widehat{\mathbf{\Sigma}}^+$.

VII. Numerical Simulation

In this section, we will study the empirical performance of both the two-step robust sparse covariance estimators and the one-step estimators, which includes the proposed estimator $\widehat{\mathbf{\Sigma}}^+$. For the two-step estimators, the first step is a robustification procedure that results in a robust pilot estimator, and the second step is a thresholding procedure that induces sparsity. To be specific, let $\widehat{\mathbf{\Sigma}}_\alpha$ denote the robust pilot estimator in a two-step estimator, where α is the robustification parameter. Let $\widehat{\mathbf{\Gamma}}_\theta(\cdot)$ denote the thresholding operator with parameter(s) θ that acts on $\widehat{\mathbf{\Sigma}}_\alpha$ and produces the final estimator $\widehat{\mathbf{\Sigma}}^f = \widehat{\mathbf{\Gamma}}_\theta(\widehat{\mathbf{\Sigma}}_\alpha)$. One example of the thresholding operator can be the positive-definite ℓ_1-penalized projection operator, which is defined as

$$\widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\mathbf{\Omega}) = \arg\min_{\mathbf{\Sigma}} \frac{1}{2} \|\mathbf{\Sigma} - \mathbf{\Omega}\|_{\text{F}}^2 - \tau \log\det \mathbf{\Sigma} + \lambda \|\mathbf{\Sigma}\|_{1,\text{off}} \quad (13)$$

It is clear for the example in (13) that, given $(\lambda, \tau) = (0, 0)$, we have $\widehat{\mathbf{\Gamma}}_{0,0}(\widehat{\mathbf{\Sigma}}_\alpha) = \widehat{\mathbf{\Sigma}}_\alpha$; Other examples include the hard/soft thresholding operators, where $\widehat{\mathbf{\Gamma}}_0(\widehat{\mathbf{\Sigma}}_\alpha) = \widehat{\mathbf{\Sigma}}_\alpha$ always

hold. In general, we can assume that $\widehat{\mathbf{\Gamma}}_\theta(\mathbf{\Omega})$ is reasonably designed such that the gap between $\mathbf{\Omega}$ and $\widehat{\mathbf{\Gamma}}_\theta(\mathbf{\Omega})$ vanishes when $\theta = 0$.

In existing literature, the tuning of two-step estimators is also broken down into two steps, where the first step is to tune the robustification parameters that dominates the behavior of the robust pilot estimator, and the second step is to tune the thresholding parameter λ that controls the sparsity in the final estimator [4], [29]. Other parameters of $\widehat{\mathbf{\Gamma}}_\theta(\cdot)$, like τ in (13), are set to be some sufficiently small number. We will follow this selection of $\tau = 10^{-6}$ [16].

Here is an example of the tuning of a two-step estimator using a V-fold Cross Validation (CV): Assuming we are tuning this estimator for a d dimensional dataset with n samples. In the first step, assume that the data follows some given distribution like a t-distribution with five degrees of freedom, and conservatively select α_1^* and α_2^* to optimize the performance of $\widehat{\mathbf{\Sigma}}_\alpha$ [29] when there are $\frac{(V-1)n}{V}$ samples and $\frac{n}{V}$ samples, respectively. In the second step, select λ by a V-fold Cross Validation in the following sense:

$$\begin{aligned}\lambda^* &= \arg\min_{\lambda\geq 0} \frac{1}{V}\sum_{v=1}^{V}\left\|\widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\widehat{\mathbf{\Sigma}}_{\alpha_1^*}^{(-v)}) - \widehat{\mathbf{\Sigma}}_{\alpha_2^*}^{(v)}\right\|_{\mathrm{F}}^2 \\ &= \arg\min_{\lambda\geq 0} \frac{1}{V}\sum_{v=1}^{V}\left\|\widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\widehat{\mathbf{\Sigma}}_{\alpha_1^*}^{(-v)}) - \widehat{\mathbf{\Gamma}}_{0,0}(\widehat{\mathbf{\Sigma}}_{\alpha_2^*}^{(v)})\right\|_{\mathrm{F}}^2\end{aligned} \tag{14}$$

where $\widehat{\mathbf{\Sigma}}_{\alpha_2^*}^{(v)}$ denotes the robust pilot estimator based on samples in the v'th fold, which includes $\frac{n}{V}$ samples, and $\widehat{\mathbf{\Sigma}}_{\alpha_1^*}^{(-v)}$ denotes the robust pilot estimator based on the samples excluding the v'th fold, which includes $\frac{(V-1)n}{V}$ samples.

The tuning of one-step estimators, however, cannot be naturally separated into two steps. Hence it is not clear how to select parameters for one-step estimators with the above method for two-step estimators. Therefore, we propose the following tuning scheme for those estimators: Let $\widehat{\mathbf{\Sigma}}_{\alpha,\lambda,\tau}$ denote any one-step robust sparse covariance estimator. In the first step, we conservatively assume that the data follows some given distribution like a t-distribution with five degrees of freedom, and select α_1^* and α_2^* to optimize the performance of $\widehat{\mathbf{\Sigma}}_{\alpha,0,0}$ in this scenario. That is, letting $\lambda = \tau = 0$ in $\widehat{\mathbf{\Sigma}}_{\alpha,\lambda,\tau}$, then focus on the tuning of α. In the second step, we select λ by a V-fold Cross Validation (CV) in the following sense:

$$\lambda^* = \arg\min_{\lambda\geq 0} \frac{1}{V}\sum_{v=1}^{V}\left\|\widehat{\mathbf{\Sigma}}_{\alpha_1^*,\lambda,\tau}^{(-v)} - \widehat{\mathbf{\Sigma}}_{\alpha_2^*,0,0}^{(v)}\right\|_{\mathrm{F}}^2 . \tag{15}$$

where $\tau = 10^{-6}$ is fixed as we mentioned earlier. This method is a generalization to the tuning of two-step estimators: If we ignore the two-step structure in (14) and substitute $\widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\widehat{\mathbf{\Sigma}}_{\alpha_i^*})$ with $\widehat{\mathbf{\Sigma}}_{\alpha_i^*,\lambda,\tau}$, $i = 1, 2$, then it coincides with the proposed tuning scheme.

We will mostly consider the two-step robust sparse convariance estimators, which includes the quantile-based estimator in [41], denoted as $\widehat{\mathbf{\Sigma}}^{\mathrm{Q}}$; the adaptive Huber's M-estimator proposed in [29], denoted as $\widehat{\mathbf{\Sigma}}^{\mathrm{M}}$; and the Huber's M-estimator utilizing pairwise-difference approach in the robustification step (3) as well as $\widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\cdot)$ in the thresholding step, denoted as $\widehat{\mathbf{\Sigma}}^{\mathrm{P}} := \widehat{\mathbf{\Gamma}}_{\lambda,\tau}(\widehat{\mathbf{\Sigma}}^{\mathrm{H}})$. Our proposed estimator (5) is the only one-step robust sparse covariance estimator so far, denoted as $\widehat{\mathbf{\Sigma}}^{+}$. The robustification parameters for all estimators are conservatively chosen to be those that would be optimal if the true distribution is a t distribution with five degrees of freedom. And the thresholding parameter λ is selected using (15).

We will compare the performance of $\widehat{\mathbf{\Sigma}}^{\mathrm{Q}}$ and $\widehat{\mathbf{\Sigma}}^{\mathrm{M}}$ with $\widehat{\mathbf{\Sigma}}^{+}$ because they are benchmark estimators for robust sparse covariance estimation. The reason we compare $\widehat{\mathbf{\Sigma}}^{\mathrm{P}}$ with the proposed estimator $\widehat{\mathbf{\Sigma}}^{+}$ here is that, we want to rule out the effect of pairwise-difference approach, just to find out exactly how much difference can be made when we switch from a two-step estimator to a one-step estimator.

Table I
QUANTITATIVE COMPARISON AMONG SEVEN DIFFERENT METHODS FOR THE BANDED SETTING

		Robust	Robust sparse				
	SCM	MoM?	Junwei Lu	Avella-Medina	PairwiseHuber_thresholding	HuberL1 (prop.)	PDHuberL1 (prop.)
$d = 100,\ n = 50$							
$\|\cdot\|_{\mathrm{F}}$	48.8570 (642.1970)	31.9230 (85.6455)	16.0938 (2.2311)	14.4988 (0.8741)	13.9719 (1.3823)	13.7575 (0.8593)	0 (0)
$\|\cdot\|_2$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
FPR	NA	NA	0.1900 (0.0078)	0.1737 (0.0015)	0.1679 (0.0017)	0.1726 (0.0008)	0 (0)
TPR	NA	NA	0.7871 (0.0037)	0.8038 (0.0013)	0.8137 (0.0012)	0.8217 (0.0009)	0 (0)
PD	0/100	0/100	100/100	0/100	91/100	86/100	0/0
Time	–	–	–	–	–	–	–
$d = 200,\ n = 50$							
$\|\cdot\|_{\mathrm{F}}$							
$\|\cdot\|_2$							
FPR							
TPR							
PD	0/100		2/100	15/100	0/100	0/100	
Time	–		–	–	–	–	

Table II
QUANTITATIVE COMPARISON AMONG SEVEN DIFFERENT METHODS FOR THE BANDED SETTING

		Robust	Robust sparse				
	SCM	MoM?	Junwei Lu	Avella-Medina	PairwiseHuber_thresholding	HuberL1 (prop.)	PDHuberL1 (prop.)
$d = 100,\ n = 50$							
$\|\cdot\|_F$	48.8570 (642.1970)	31.9230 (85.6455)	16.0938 (2.2311)	14.4988 (0.8741)	13.9719 (1.3823)	13.7575 (0.8593)	0 (0)
$\|\cdot\|_2$	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
FPR	NA	NA	0.1900 (0.0078)	0.1737 (0.0015)	0.1679 (0.0017)	0.1726 (0.0008)	0 (0)
TPR	NA	NA	0.7871 (0.0037)	0.8038 (0.0013)	0.8137 (0.0012)	0.8217 (0.0009)	0 (0)
PD	0/100	0/100	100/100	0/100	91/100	86/100	0/0
Time	–	–	–	–	–	–	–
$d = 200,\ n = 50$							
$\|\cdot\|_F$							
$\|\cdot\|_2$							
FPR							
TPR							
PD	0/100		2/100	15/100	0/100	0/100	
Time	–		–	–	–	–	

VIII. Conclusion

XX

XX

XX

XX

XX

Appendix A
Proofs of Statistical Theory

In this appendix, we first provide some necessary technical lemmata, and then provide the proofs of all the statistical theoretical results in Section V and Section VI.

A. Technical Lemmata

In this subsection, we derive some technical lemmata by considering

$$\widetilde{\boldsymbol{\Sigma}} \in \arg\min_{\boldsymbol{\Sigma} \succ \epsilon \boldsymbol{I}} \left\{ L_\alpha(\boldsymbol{\Sigma}) - \tau \log\det\left(\boldsymbol{\Sigma} - \epsilon \boldsymbol{I}\right) + \lambda \left\|\boldsymbol{\Sigma}\right\|_{1,\text{off}} \right\}, \tag{16}$$

where $\epsilon \geq 0$ lower bounds the minimal singular value of the final estimation $\widetilde{\boldsymbol{\Sigma}}$. This definition is a slight generalization of Problem (5).

Lemma 11. *For any positive-definite $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ satisfying $\boldsymbol{\Sigma}_{\overline{\mathcal{S}}} = \boldsymbol{0}$, provided $\boldsymbol{\Sigma} \succ \epsilon \boldsymbol{I}$ and $\lambda > \left\|\nabla L_\alpha(\boldsymbol{\Sigma})\right\|_\infty + \tau \left\|\left(\boldsymbol{\Sigma} - \epsilon \boldsymbol{I}\right)^{-1}\right\|_\infty$, we have*

$$\begin{aligned}
&\left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 \\
\leq &\left(\lambda - \left\|\nabla L_\alpha(\boldsymbol{\Sigma})\right\|_\infty - \tau \left\|\left(\boldsymbol{\Sigma} - \epsilon \boldsymbol{I}\right)^{-1}\right\|_\infty\right)^{-1} \\
&\cdot \left(\lambda + \left\|\nabla L_\alpha(\boldsymbol{\Sigma})\right\|_\infty + \tau \left\|\left(\boldsymbol{\Sigma} - \epsilon \boldsymbol{I}\right)^{-1}\right\|_\infty\right) \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1
\end{aligned}$$

Proof: Let $f(\boldsymbol{\Sigma}) \coloneqq L_\alpha(\boldsymbol{\Sigma}) - \tau \log\det\left(\boldsymbol{\Sigma} - \epsilon \boldsymbol{I}\right)$. For $\boldsymbol{\Xi} \in \partial \left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$, define

$$\boldsymbol{U}(\boldsymbol{\Xi}) = \nabla f(\widetilde{\boldsymbol{\Sigma}}) + \lambda \boldsymbol{\Xi} \in \mathbb{R}^{d\times d}.$$

Optimality condition of (5) implies $\inf_{\boldsymbol{\Xi}} \boldsymbol{U}(\boldsymbol{\Xi}) = 0$. By convexity of $f(\boldsymbol{\Sigma})$:

$$\langle \nabla f(\widetilde{\boldsymbol{\Sigma}}) - \nabla f(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \geq 0.$$

Therefore,

$$\begin{aligned}
&\left\|\boldsymbol{U}(\boldsymbol{\Xi})\right\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
\geq &\langle \boldsymbol{U}(\boldsymbol{\Xi}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
= &\langle \nabla f(\widetilde{\boldsymbol{\Sigma}}) - \nabla f(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
&+ \langle \nabla f(\boldsymbol{\Sigma}), \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle + \langle \lambda \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
\geq &0 - \left\|\nabla f(\boldsymbol{\Sigma})\right\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 + \langle \lambda \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\langle \lambda \boldsymbol{\Xi}, \widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \rangle \\
= &\lambda \langle \boldsymbol{\Xi}_{\overline{\mathcal{S}}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}} \rangle + \lambda \langle \boldsymbol{\Xi}_{\mathcal{S}}, (\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}} \rangle \\
\geq &\lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_{1,\text{off}}
\end{aligned}$$

Together, the last two displays imply

$$\begin{aligned}
&\left\|\boldsymbol{U}(\boldsymbol{\Xi})\right\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
\geq &- \left\|\nabla f(\boldsymbol{\Sigma})\right\|_\infty \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 \\
&+ \lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1
\end{aligned}$$

Since the right-hand side of this inequality does not depend on $\boldsymbol{\Xi}$, taking the infimum with respect to $\boldsymbol{\Xi} \in \partial \left\|\widetilde{\boldsymbol{\Sigma}}\right\|_{1,\text{off}}$ on both sides to reach

$$\begin{aligned}
0 \geq &- \left\|\nabla f(\boldsymbol{\Sigma})\right\|_\infty \left\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_1 \\
&+ \lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\overline{\mathcal{S}}}\right\|_1 - \lambda \left\|(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})_{\mathcal{S}}\right\|_1
\end{aligned}$$

Decompose $\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}\right\|_1$ as $\left\|(\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma})_{\mathcal{S}}\right\|_1+\left\|(\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma})_{\overline{\mathcal{S}}}\right\|_1$, the stated result follows immediately. ∎

Lemma 12. *Let $\mathbf{\Sigma}^* \succ \epsilon \boldsymbol{I}$. Assume $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau\left\|(\mathbf{\Sigma}^*-\epsilon\boldsymbol{I})^{-1}\right\|_\infty \le \lambda/2$, then any solution $\widetilde{\mathbf{\Sigma}}$ to (5) satisfies $\widetilde{\mathbf{\Sigma}} \in \mathbf{\Sigma}^* + \mathbb{C}(4s^{1/2})$, with*

$$\mathbb{C}(l) \coloneqq \left\{\mathbf{\Delta} \in \mathbb{R}^{d\times d} : \|\mathbf{\Delta}\|_1 \le l\|\mathbf{\Delta}\|_F\right\}.$$

Moreover, assume $C\alpha > (3/2)\cdot\kappa^{-1}\lambda s^{1/2}$ for constant C. Then, condition on the event

$$\mathcal{E}_1(C\alpha,\kappa)\cap\left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau\left\|(\mathbf{\Sigma}^*-\epsilon\boldsymbol{I})^{-1}\right\|_\infty \le \lambda/2\right\}, \tag{17}$$

we have

$$\begin{aligned}\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_F &\le \kappa^{-1}\Bigg\{\lambda s^{1/2} + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_F \\ &\qquad +\tau\left\|\left((\mathbf{\Sigma}^*-\epsilon\boldsymbol{I})^{-1}\right)_{\mathcal{S}}\right\|_{\mathrm{F}}\Bigg\} \\ &\le (3/2)\cdot\kappa^{-1}\lambda s^{1/2}.\end{aligned}$$

Proof: Condition on the stated event, Lemma 11 indicates

$$\left\|(\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1 \le 3\left\|(\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_1.$$

Therefore,

$$\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_1 \le 4s^{1/2}\left\|\widetilde{\mathbf{\Sigma}}-\mathbf{\Sigma}^*\right\|_{\mathrm{F}},$$

that is, $\widetilde{\mathbf{\Sigma}} \in \mathbf{\Sigma}^* + \mathbb{C}(4s^{1/2})$.

Now we prove the second statement. Define $\eta = \sup\left\{u\in[0,1] : (1-u)\mathbf{\Sigma}^* + u\widetilde{\mathbf{\Sigma}} \in \mathbb{B}^\infty(C\alpha)\right\}$. Note that $\eta = 1$ if $\widetilde{\mathbf{\Sigma}} \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$ and $\eta \in (0,1)$ otherwise.

Let $\widetilde{\mathbf{\Sigma}}_\eta \coloneqq (1-\eta)\mathbf{\Sigma}^* + \eta\widetilde{\mathbf{\Sigma}}$. Notice that if $\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_\infty < C\alpha$, then $\widetilde{\mathbf{\Sigma}}_\eta = \widetilde{\mathbf{\Sigma}}$. By the convexity of Huber loss, we have

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}_\eta) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\rangle \\ &\qquad \le \eta\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle\end{aligned} \tag{18}$$

Since $\widetilde{\mathbf{\Sigma}}_\eta \in \mathbf{\Sigma}^* + \mathbb{B}^\infty(C\alpha)$, and condition on event $\mathcal{E}_1(C\alpha,\kappa)$, we have

$$\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}_\eta) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\rangle \ge \kappa\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 \tag{19}$$

Write $f(\mathbf{\Sigma}) \coloneqq L_\alpha(\mathbf{\Sigma}) - \tau\log\det(\mathbf{\Sigma}-\epsilon\boldsymbol{I})$. Now we upper bound the right-hand side of (18). For $\mathbf{\Xi} \in \partial\left\|\widetilde{\mathbf{\Sigma}}\right\|_{1,\text{off}}$, write

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle \\ &= \underbrace{\langle\boldsymbol{U}(\mathbf{\Xi}), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle}_{\coloneqq\Pi_1} - \underbrace{\langle\nabla f(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle}_{\coloneqq\Pi_2} \\ &\quad - \underbrace{\langle\lambda\mathbf{\Xi}, \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle}_{\coloneqq\Pi_3} - \underbrace{\tau\langle(\mathbf{\Sigma}^*)^{-1} - \widetilde{\mathbf{\Sigma}}^{-1}, \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle}_{\ge 0}\end{aligned} \tag{20}$$

where $\boldsymbol{U}(\mathbf{\Xi}) \coloneqq \nabla f(\widetilde{\mathbf{\Sigma}}) + \lambda\mathbf{\Xi} \in \mathbb{R}^{d\times d}$. We have

$$\begin{aligned}|\Pi_1| &\le \|(\boldsymbol{U}(\mathbf{\Xi}))_{\mathcal{S}}\|_{\mathrm{F}}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + \|\boldsymbol{U}(\mathbf{\Xi})\|_\infty\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1 \\ |\Pi_2| &\le \|\nabla f(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\ &\qquad + \|\nabla f(\mathbf{\Sigma}^*)\|_\infty\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1.\end{aligned}$$

Turning to Π_3, decompose $\lambda\mathbf{\Xi}$ and $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$ according to $\mathcal{S}\cup\overline{\mathcal{S}}$ to reach

$$\Pi_3 = \langle(\lambda\mathbf{\Xi})_{\mathcal{S}}, (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\mathcal{S}}\rangle + \langle(\lambda\mathbf{\Xi})_{\overline{\mathcal{S}}}, (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\rangle$$

Since $\mathbf{\Sigma}^*_{\overline{\mathcal{S}}} = \mathbf{0}$ and $\mathbf{\Xi} \in \partial\left\|\widetilde{\mathbf{\Sigma}}\right\|_{1,\text{off}}$, we have

$$\begin{aligned}\langle(\lambda\mathbf{\Xi})_{\overline{\mathcal{S}}}, (\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\rangle &= \langle(\lambda\mathbf{\Xi})_{\overline{\mathcal{S}}}, \widetilde{\mathbf{\Sigma}}_{\overline{\mathcal{S}}}\rangle \\ &= \lambda\left\|\widetilde{\mathbf{\Sigma}}_{\overline{\mathcal{S}}}\right\|_1 = \lambda\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1.\end{aligned}$$

Therefore,

$$\Pi_3 \ge \lambda\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1 - \lambda s^{1/2}\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\mathcal{S}}\right\|_{\mathrm{F}}.$$

Combining (20) with our estimation for Π_1, Π_2 and Π_3, we have

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle \\ &\le -\left\{\lambda - \|\nabla f(\mathbf{\Sigma}^*)\|_\infty - \|\boldsymbol{U}(\mathbf{\Xi})\|_\infty\right\}\cdot\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1 \\ &\quad + \|\nabla f(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + \|(\boldsymbol{U}(\mathbf{\Xi}))_{\mathcal{S}}\|_{\mathrm{F}}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\ &\quad + \lambda s^{1/2}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}\end{aligned}$$

Taking the infimum with respect to $\mathbf{\Xi} \in \partial\left\|\widetilde{\mathbf{\Sigma}}\right\|_{1,\text{off}}$ on both sides, it follows that

$$\begin{aligned}&\langle\nabla L_\alpha(\widetilde{\mathbf{\Sigma}}) - \nabla L_\alpha(\mathbf{\Sigma}^*), \widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\rangle \\ &\le \left(\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \left\|\tau\left(\mathbf{\Sigma}^* - \epsilon\boldsymbol{I}\right)^{-1}\right\|_\infty - \lambda\right)\left\|(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)_{\overline{\mathcal{S}}}\right\|_1 \\ &\quad + \|\nabla f(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}}\cdot\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} + \lambda s^{1/2}\left\|\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}.\end{aligned} \tag{21}$$

With $\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^* = \eta(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*)$, it follows from (18), (19) and (21) that condition on the stated event (17),

$$\begin{aligned}&\kappa\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_{\mathrm{F}}^2 \le \\ &\left\{\lambda s^{1/2} + \|\nabla f(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}}\right\}\cdot\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_{\mathrm{F}},\end{aligned}$$

which implies that

$$\begin{aligned}&\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \\ &\le \kappa^{-1}\left\{\lambda s^{1/2} + \|\nabla L_\alpha(\mathbf{\Sigma}^*)_{\mathcal{S}}\|_{\mathrm{F}} + \tau\left\|\left((\mathbf{\Sigma}^* - \epsilon\boldsymbol{I})^{-1}\right)_{\mathcal{S}}\right\|_{\mathrm{F}}\right\} \\ &\le \kappa^{-1}\{\lambda s^{1/2} + \frac{1}{2}\lambda s^{1/2}\} = \frac{3}{2}\kappa^{-1}\lambda s^{1/2} < C\alpha.\end{aligned} \tag{22}$$

Since $\left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_\infty \le \left\|\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} < C\alpha$, $\widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^*$ falls in the interior of $\mathbb{B}^\infty(C\alpha)$. Hence $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^* = \widetilde{\mathbf{\Sigma}}_\eta - \mathbf{\Sigma}^* \in \mathbb{B}^\infty(C\alpha)$. Consequently, (22) also holds for $\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^*$. ∎

B. Proof of Proposition 6

Proof: $\left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \le 3\lambda s^{1/2}$ follows immediately from Lemma 12 with $\epsilon = 0$ and $C = \kappa = 1/2$. Combining this with $\widehat{\mathbf{\Sigma}}^+ \in \mathbf{\Sigma}^* + \mathbb{C}(4s^{1/2})$, yields $\left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_1 \le 12\lambda s$. ∎

C. Proof of Proposition 7

Proof: Recall $N = n(n-1)/2$. For fixed $k, l \in [d] \times [d]$, let $z_{i,j} = (x_{ik} - x_{jk})(x_{il} - x_{jl})/2$ and define

$$D_{kl} = \frac{1}{N} \sum_{1 \le i < j \le n} 1\left(|\Sigma^*_{kl} - z_{i,j}| \le \alpha/2\right). \tag{23}$$

By Chebyshev's inequality,

$$\mathbb{E}[D_{kl}] = \Pr\left(|\Sigma^*_{kl} - z_{i,j}| \le \alpha/2\right) \ge 1 - 4K/\alpha^2 > (1+\kappa)/2.$$

The last inequality holds because $4K/\alpha^2 < (1-\kappa)/2$, which follows from $\alpha \asymp \sqrt{Kn/\log d}$ and by taking $n \gtrsim \log d$.

For each fixed $k, l \in [d]$, let $h(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1(|\Sigma^*_{kl} - z_{i,j}| \le \alpha/2)$. Note that the right hand side of (23) is a U-statistic with a bounded kernel of order two. With Hoeffding's inequality for U-statistics [47],

$$\Pr\left(|D_{kl} - \mathbb{E}[D_{kl}]| > \sqrt{\frac{\log(2/\delta)}{2\lfloor n/2 \rfloor}}\right) \le \delta.$$

Taking $\delta = 2 \cdot \exp\left(-(1-\kappa)^2 n/4\right)$ yields

$$\Pr\left(|D_{kl} - \mathbb{E}[D_{kl}]| > \frac{1-\kappa}{2}\right) \le 2 \cdot \exp\left(-(1-\kappa)^2 n/4\right).$$

Therefore, D_{kl} is concentrated around its mean, which implies

$$\begin{aligned}
&\Pr\left\{D_{kl} < \kappa\right\} \\
\le &\Pr\left\{|D_{kl} - \mathbb{E}[D_{kl}]| \ge (1-\kappa)/2\right\} \\
\le &2 \cdot \exp\left(-(1-\kappa)^2 n/4\right).
\end{aligned}$$

With union bound we have

$$\Pr\left[\min_{k,l} D_{kl} < \kappa\right] \le 2d^2 \cdot \exp\left(-(1-\kappa)^2 n/4\right) < 2/d,$$

where the last inequality follows by taking $n \ge 12 \log d/(1-\kappa)^2$. Let $\mathcal{G}_{kl} := \{(i,j) : i < j \text{ and } |\Sigma^*_{kl} - z_{i,j}| \le \alpha/2\}$. Under the event that $\min_{k,l} D_{kl} \ge \kappa$,

$$\begin{aligned}
&\frac{1}{N} \sum_{1 \le i < j \le n} \{\rho'_\alpha(\Sigma_{1,kl} - z_{i,j}) - \rho'_\alpha(\Sigma_{2,kl} - z_{i,j})\} \\
&\qquad \cdot (\Sigma_{1,kl} - \Sigma_{2,kl}) \\
\ge &\frac{1}{N} \sum_{(i,j) \in \mathcal{G}_{kl}} \{\rho'_\alpha(\Sigma_{1,kl} - z_{i,j}) - \rho'_\alpha(\Sigma_{2,kl} - z_{i,j})\} \\
&\qquad \cdot (\Sigma_{1,kl} - \Sigma_{2,kl}) \\
= &\frac{1}{N} \sum_{(i,j) \in \mathcal{G}_{kl}} (\Sigma_{1,kl} - \Sigma_{2,kl})^2 \\
\ge &\kappa (\Sigma_{1,kl} - \Sigma_{2,kl})^2
\end{aligned}$$

The second last equality holds since $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathbb{B}^\infty(\alpha/2)$ implies $|\Sigma_{1,kl} - z_{i,j}| \le \alpha/2$ and $|\Sigma_{2,kl} - z_{i,j}| \le \alpha/2$ for $(i,j) \in \mathcal{G}_{kl}$. The last inequality follows from $|\mathcal{G}_{kl}|/N = D_{kl}$. Therefore

$$\begin{aligned}
&\langle \nabla L_\alpha(\boldsymbol{\Sigma}_1) - \nabla L_\alpha(\boldsymbol{\Sigma}_2), \boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2 \rangle \\
= &\sum_{k,l} \frac{1}{N} \sum_{1 \le i < j \le n} \{\rho'_\alpha(\Sigma_{1,kl} - z_{i,j}) - \rho'_\alpha(\Sigma_{2,kl} - z_{i,j})\} \\
&\qquad \cdot (\Sigma_{1,kl} - \Sigma_{2,kl}) \\
\ge &\kappa \cdot \|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\|_{\mathrm{F}}^2
\end{aligned}$$

with at least $1 - 2/d$ probability. ∎

D. Proof of Lemma 8

Proof: We adopt the following notations:

$$\begin{aligned}
\mathbf{B}^* &:= \mathbb{E}[\nabla L_\alpha(\boldsymbol{\Sigma}^*)] \\
\mathbf{W}^* &:= \nabla L_\alpha(\boldsymbol{\Sigma}^*) - \mathbb{E}[\nabla L_\alpha(\boldsymbol{\Sigma}^*)].
\end{aligned}$$

We first analyze $\mathbf{B}^*$. For each $k, l \in [d]$, let $\epsilon_{kl} := \Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2$, then

$$\begin{aligned}
|\mathbb{E}[\rho'_\alpha(\epsilon_{kl})]| &= |\mathbb{E}[\epsilon_{kl} I(|\epsilon_{kl}| \le \alpha) + \alpha \mathrm{sgn}(\epsilon_{kl}) I(|\epsilon_{kl}| > \alpha)]| \\
&= |\mathbb{E}[\epsilon_{kl} + (\alpha \mathrm{sign}(\epsilon_{kl}) - \epsilon_{kl}) I(|\epsilon_{kl}| > \alpha)]| \\
&= |\mathbb{E}\{[\epsilon_{kl} - \alpha \mathrm{sign}(\epsilon_{kl})] I(|\epsilon_{kl}| > \alpha)\}| \\
&\le |\mathbb{E}[(|\epsilon_{kl}| - \alpha \mathrm{sign}(\epsilon_{kl})) I(|\epsilon_{kl}| > \alpha)]| \\
&\le \frac{\left|\mathbb{E}[(\epsilon_{kl}^2 - \alpha^2) I(|\epsilon_{kl}| > \alpha)]\right|}{\alpha} \\
&< \frac{K}{\alpha}.
\end{aligned}$$

hence we have $|(\mathbf{B}^*)_{kl}| = |\mathbb{E}[\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2)]| < \frac{K}{\alpha}$. Next, we will analyze $\mathbf{W}^*$. By definition,

$$\begin{aligned}
W^*_{kl} = \frac{1}{N} \sum_{1 \le i < j \le n} &\{\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2) \\
&- \mathbb{E}\left[\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2)\right]\}.
\end{aligned}$$

Given that $|\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2)| \le \alpha$, for all $m \ge 2$:

$$\begin{aligned}
&\mathbb{E}[\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2)]^m \\
\le &\alpha^{m-2} \cdot \mathrm{Var}[\rho'_\alpha(\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2)] \\
\le &\alpha^{m-2} \cdot \mathrm{Var}[\Sigma^*_{k\ell} - y_{ik} y_{i\ell}/2] \\
\le &\alpha^{m-2} K \le \alpha^{m-2} K \cdot m!/2
\end{aligned}$$

The second inequality follows given $\rho'_\alpha(\cdot)$ is 1-Lipschitz. With Bernstein's inequality for U-statistics [47], for any $t \ge 0$,

$$\Pr\left(|W^*_{kl}| \ge t\right) \le 2 \exp\left(\frac{-\lfloor n/2 \rfloor \cdot t^2}{2\left(K + \alpha t/3\right)}\right)$$

By taking $t = \sqrt{12K \log d/n} + 4\alpha \log d/n$,

$$\begin{aligned}
&\Pr\left(|W^*_{kl}| \ge \sqrt{\frac{12K \log d}{n}} + \alpha \frac{4 \log d}{n}\right) \\
&\le 2 \exp\left(\frac{-\left(\sqrt{12Kn \log d} + 4\alpha \cdot \log d\right)^2}{4\left(Kn + (\alpha/3) \cdot \sqrt{12Kn \log d} + (4\alpha^2/3) \log d\right)}\right) \\
&\le 2 \exp\Bigg(-\frac{4Kn + (8\alpha/3) \cdot \sqrt{12Kn \log d} + \left(16\alpha^2/3\right) \log d}{4Kn + (4\alpha/3) \cdot \sqrt{12Kn \log d} + (16\alpha^2/3) \log d} \\
&\qquad \cdot 3 \log d\Bigg) \\
&< \frac{2}{d^3}
\end{aligned}$$

In conjunction with the union bound,

$$\Pr\left(\|\mathbf{W}^*\|_\infty \ge \sqrt{12K \log d/n} + 4\alpha \log d/n\right) < \frac{2}{d}.$$

Recall that $\|\mathbf{B}^*\|_\infty < K/\alpha$. With $\nabla L_\alpha(\boldsymbol{\Sigma}^*) = \mathbf{B}^* + \mathbf{W}^*$,

$$\|\nabla L_\alpha(\boldsymbol{\Sigma}^*)\|_\infty < \sqrt{12K \log d/n} + 4\alpha \log d/n + K/\alpha \tag{24}$$

holds with at least $1-2/d$ probability.

Given $\alpha \asymp \sqrt{Kn/\log d}$, from (24) we have $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty \lesssim \sqrt{K\log d/n}$. By taking $\lambda \asymp \sqrt{K\log d/n}$, $\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2$ holds with at least $1-2/d$ probability. ∎

E. Proof of Theorem 9

Proof: The proof combines Proposition 6 with Proposition 7 and Lemma 8. By Lemma 8, event $\left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2\right\}$ happens with at least $1-2/d$ probability.

With $n \gtrsim s^{1/2}\log d$, we have $\alpha > 6\lambda s^{1/2}$. Proposition 7 indicates that $\mathcal{E}_1(\alpha/2, 1/2)$ happens with at least $1-2/d$ probability. With union bound, event $\mathcal{E}_1(\alpha/2,1/2)\cap\left\{\|\nabla L_\alpha(\mathbf{\Sigma}^*)\|_\infty + \tau\left\|(\mathbf{\Sigma}^*)^{-1}\right\|_\infty \leq \lambda/2\right\}$ holds with at least $1-4/d$ probability. Under this event and by Proposition 6,

$$\left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \leq 3\lambda s^{1/2} \quad \text{and} \quad \left\|\widehat{\mathbf{\Sigma}}^+ - \mathbf{\Sigma}^*\right\|_1 \leq 12\lambda s.$$

Then it suffices to recall $\lambda \asymp \sqrt{K\log d/n}$. ∎

F. Proof of Theorem 10

Proof: Consider using the log-determinant barrier method to solve problem (12)

$$\min_{\mathbf{\Sigma}\succ\epsilon\boldsymbol{I}} \{tg(\mathbf{\Sigma}) - \log\det(\mathbf{\Sigma}-\epsilon\boldsymbol{I})\}$$

where $g(\mathbf{\Sigma}) \coloneqq L_\alpha(\mathbf{\Sigma}) + \lambda\|\mathbf{\Sigma}\|_{1,\text{off}}$ and $t=\tau^{-1}$. Equivalently, define

$$\widehat{\mathbf{\Sigma}}_\tau^+ \in \arg\min_{\mathbf{\Sigma}\succ\epsilon\boldsymbol{I}} \{g(\mathbf{\Sigma}) - \tau\log\det(\mathbf{\Sigma}-\epsilon\boldsymbol{I})\}. \tag{25}$$

The intuition of the following proof is to take the limit of $\widehat{\mathbf{\Sigma}}_\tau^+$ as $\tau\to 0$. Using the optimality condition for problem (25), we have

$$\nabla L_\alpha(\widehat{\mathbf{\Sigma}}_\tau^+) + \lambda\mathbf{\Xi} - \tau\left(\widehat{\mathbf{\Sigma}}_\tau^+ - \epsilon\boldsymbol{I}\right)^{-1} = \mathbf{0}$$

with some $\mathbf{\Xi}\in\partial\left\|\widehat{\mathbf{\Sigma}}_\tau^+\right\|_{1,\text{off}}$. Consider

$$\min_{\mathbf{\Sigma}\succeq\epsilon\boldsymbol{I}} g(\mathbf{\Sigma}) - \tau\left\langle\left(\widehat{\mathbf{\Sigma}}_\tau^+ - \epsilon\boldsymbol{I}\right)^{-1}, \mathbf{\Sigma}\right\rangle,$$

we can see that $\widehat{\mathbf{\Sigma}}_\tau^+$ minimizes this problem. Further, for all $\mathbf{\Sigma}\succeq\mathbf{0}$, we have $\left\langle\left(\widehat{\mathbf{\Sigma}}_\tau^+ - \epsilon\boldsymbol{I}\right)^{-1}, \mathbf{\Sigma}\right\rangle \geq 0$ by the property of the positive-semidefinite cone. Hence

$$\begin{aligned}\min_{\mathbf{\Sigma}\succeq\mathbf{0}} g(\mathbf{\Sigma}) \geq& \min_{\mathbf{\Sigma}\succeq\mathbf{0}} g(\mathbf{\Sigma}) - \tau\left\langle\left(\widehat{\mathbf{\Sigma}}_\tau^+\right)^{-1}, \mathbf{\Sigma}\right\rangle \\ =& g(\widehat{\mathbf{\Sigma}}_\tau^+) - \tau\left\langle\left(\widehat{\mathbf{\Sigma}}_\tau^+\right)^{-1}, \widehat{\mathbf{\Sigma}}_\tau^+\right\rangle = g(\widehat{\mathbf{\Sigma}}_\tau^+) - \tau d\end{aligned}$$

Let $g^* \coloneqq \min_{\mathbf{\Sigma}\succeq\mathbf{0}} g(\mathbf{\Sigma})$. So far, we have shown that

$$g\left(\widehat{\mathbf{\Sigma}}_\tau^+\right) \leq g^* + \tau d,$$

which can also be derived following the approach in [48]. In particular, for all $\tau\leq 1$, we can see that $g\left(\widehat{\mathbf{\Sigma}}_\tau^+\right) \leq g^* + d$. That is, $\widehat{\mathbf{\Sigma}}_\tau^+ \in \{\mathbf{\Sigma}\succeq\mathbf{0} : g(\mathbf{\Sigma}) \leq g^* + d\} \coloneqq S$, with S being a compact set. Therefore, there exists decreasing sequence $\{\tau_n\}_{n=1}^\infty$ such that $\tau_n\to 0$ and $\widehat{\mathbf{\Sigma}}_{\tau_n}^+$ converges to some $\widehat{\mathbf{\Sigma}}_\epsilon \in S$. Moreover, $\widehat{\mathbf{\Sigma}}_\epsilon$is a minimizer for problem (12) because

$$g\left(\widehat{\mathbf{\Sigma}}_\epsilon\right) = \lim_{n\to\infty} g\left(\widehat{\mathbf{\Sigma}}_{\tau_n}^+\right) = g^*.$$

Meanwhile, similar to the proof of Theorem 9, we can apply Lemma 12 to $\widehat{\mathbf{\Sigma}}_\tau^+$, which concludes that with at least $1-4/d$ probability, $\left\|\widehat{\mathbf{\Sigma}}_\tau^+ - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{Ks\log d}{n}}$ and $\left\|\widehat{\mathbf{\Sigma}}_\tau^+ - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K\log d}{n}}$ hold for all τ sufficiently small. In the limit, this implies that $\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{Ks\log d}{n}}$ and $\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_1 \lesssim s\sqrt{\frac{K\log d}{n}}$ as well.

Now we prove that $\widehat{\mathbf{\Sigma}}_\epsilon$ is the unique minimizer for problem (12). We already know that $\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} \lesssim \sqrt{\frac{Ks\log d}{n}}$. By taking $n \gtrsim s^{1/2}\log d$, we have $\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_\infty \leq \left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}^*\right\|_{\mathrm{F}} < \alpha/2$.

To put it in words, the interior of $\mathbb{B}^\infty(\alpha/2)$ contains $\widehat{\mathbf{\Sigma}}_\epsilon$. The set of optimal solutions is convex. Therefore, if $\widehat{\mathbf{\Sigma}}_\epsilon$ is not the unique minimizer, there would be some $\mathbf{\Sigma}_0\in\mathbb{B}^\infty(\alpha/2)$, such that $g(\mathbf{\Sigma}_0) = g^*$ and $\mathbf{\Sigma}_0 \neq \widehat{\mathbf{\Sigma}}_\epsilon$. Using Proposition 7, we have

$$\left\langle\nabla L_\alpha(\widehat{\mathbf{\Sigma}}_\epsilon) - \nabla L_\alpha(\mathbf{\Sigma}_0), \widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\rangle \geq \frac{1}{2}\left\|\widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\|_{\mathrm{F}}^2 > 0. \tag{26}$$

However, by first-order optimality conditions,

$$\left\langle\nabla L_\alpha(\widehat{\mathbf{\Sigma}}_\epsilon) + \lambda\widehat{\mathbf{\Xi}}, \mathbf{\Sigma}_0 - \widehat{\mathbf{\Sigma}}_\epsilon\right\rangle \geq 0$$

$$\left\langle\nabla L_\alpha(\mathbf{\Sigma}_0) + \lambda\mathbf{\Xi}_0, \widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\rangle \geq 0$$

with $\widehat{\mathbf{\Xi}}\in\partial\left\|\widehat{\mathbf{\Sigma}}_\epsilon\right\|_{1,\text{off}}$ and $\mathbf{\Xi}_0\in\partial\|\mathbf{\Sigma}_0\|_{1,\text{off}}$. Therefore,

$$\left\langle\nabla L_\alpha(\widehat{\mathbf{\Sigma}}_\epsilon) - \nabla L_\alpha(\mathbf{\Sigma}_0) + \lambda\widehat{\mathbf{\Xi}} - \lambda\mathbf{\Xi}_0, \widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\rangle \leq 0$$

By convexity of the off-diagonal ℓ_1 penalty, we have $\left\langle\lambda\widehat{\mathbf{\Xi}} - \lambda\mathbf{\Xi}_0, \widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\rangle \geq 0$. Hence

$$\left\langle\nabla L_\alpha(\widehat{\mathbf{\Sigma}}_\epsilon) - \nabla L_\alpha(\mathbf{\Sigma}_0), \widehat{\mathbf{\Sigma}}_\epsilon - \mathbf{\Sigma}_0\right\rangle \leq 0,$$

which contradicts with (26). We can now conclude that $\widehat{\mathbf{\Sigma}}_\epsilon$ is the unique minimizer for problem (12). ∎

REFERENCES

[1] Q. Wei and Z. Zhao, "Large covariance matrix estimation with oracle statistical rate," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2023.

[2] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[3] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, 2005.

[4] Y. He, P. Liu, X. Zhang, and W. Zhou, "Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis," *Statistics in Medicine*, vol. 40, no. 15, pp. 3499–3515, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8979

[5] A. Serra, P. Coretto, M. Fratello, and R. Tagliaferri, "Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data," *Bioinformatics*, vol. 34, 10 2017.

[6] O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *J. Empir. Finance.*, vol. 10, no. 5, pp. 603–621, 2003.

[7] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2342–2357, 2018.

[8] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Trans. Signal Process.*, vol. 67, no. 7, pp. 1681–1695, 2019.

[9] Y. Chen, A. Wiesel, Y. C. Eldar, and A. O. Hero, "Shrinkage algorithms for MMSE covariance estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5016–5029, 2010.

[10] Y. Sun, P. Babu, and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3576–3590, 2016.

[11] A. Aubry, A. De Maio, and L. Pallotta, "A geometric approach to covariance matrix estimation and its applications to radar problems," *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 907–922, 2018.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *Ann. Statist.*, vol. 36, no. 6, pp. 2577–2604, 2008.

[14] N. El Karoui, "Operator norm consistent estimation of large-dimensional sparse covariance matrices," *Ann. Statist.*, vol. 36, no. 6, pp. 2717–2756, 2008.

[15] A. J. Rothman, E. Levina, and J. Zhu, "Generalized thresholding of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 104, no. 485, pp. 177–186, 2009.

[16] A. J. Rothman, "Positive definite estimators of large covariance matrices," *Biometrika*, vol. 99, no. 3, pp. 733–740, 2012.

[17] Y. Cui, C. Leng, and D. Sun, "Sparse estimation of high-dimensional correlation matrices," *Comput. Stat. Data Anal.*, vol. 93, pp. 390–403, 2016.

[18] L. Xue, S. Ma, and H. Zou, "Positive-definite ℓ_1-penalized estimation of large covariance matrices," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1480–1491, 2012.

[19] B. P. Lan Wang and R. Li, "A high-dimensional nonparametric multivariate test for mean vector," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1658–1669, 2015, pMID: 26848205. [Online]. Available: https://doi.org/10.1080/01621459.2014.988215

[20] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, pp. 223–236, 2001. [Online]. Available: https://doi.org/10.1080/713665670

[21] Y. Ke, S. Minsker, Z. Ren, Q. Sun, and W.-X. Zhou, "User-friendly covariance estimation for heavy-tailed distributions," *Statistical Science*, vol. 34, no. 3, pp. 454–471, 2019.

[22] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices," *Econom. J.*, vol. 19, no. 1, pp. C1–C32, 2016.

[23] O. Catoni, "Challenging the empirical mean and empirical variance: A deviation study," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 48, no. 4, pp. 1148 – 1185, 2012. [Online]. Available: https://doi.org/10.1214/11-AIHP454

[24] H. Wang and A. Ramdas, "Catoni-style confidence sequences for heavy-tailed mean estimation," *Stochastic Processes and their Applications*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246485456

[25] L. Wang, C. Zheng, W.-X. Zhou, and W.-X. Zhou, "A new principle for tuning-free huber regression," *Statistica Sinica*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:89609322

[26] O. Catoni, "Pac-bayesian bounds for the gram matrix and least squares regression with a random design," 03 2016.

[27] S. Minsker, "Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries," *The Annals of Statistics*, vol. 46, no. 6A, pp. 2871 – 2903, 2018. [Online]. Available: https://doi.org/10.1214/17-AOS1642

[28] S. Minsker and X. Wei, "Robust modifications of U-statistics and applications to covariance estimation problems," *Bernoulli*, vol. 26, no. 1, pp. 694 – 727, 2020. [Online]. Available: https://doi.org/10.3150/19-BEJ1149

[29] M. Avella-Medina, H. S. Battey, J. Fan, and Q. Li, "Robust estimation of high-dimensional covariance and precision matrices." *Biometrika*, vol. 105 2, pp. 271–284, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49237226

[30] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, pp. 492–518, 1964. [Online]. Available: https://api.semanticscholar.org/CorpusID:121252793

[31] Q. Sun, W. Zhou, and J. Fan, "Adaptive huber regression," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 254–265, Jan. 2020, publisher Copyright: © 2019, © 2019 American Statistical Association.

[32] J. Fan, Q. Li, and Y. Wang, "Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 79, no. 1, pp. 247–265, 2017. [Online]. Available: http://www.jstor.org/stable/44681770

[33] A. S. Nemirovsky and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, ser. A Wiley-Interscience publication. Wiley, 1983. [Online]. Available: https://mathscinet.ams.org/mathscinet/relay-station?mr=0702836

[34] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, "Sub-gaussian mean estimators," *arXiv: Statistics Theory*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:26805883

[35] S. Minsker and N. Strawn, "Distributed statistical estimation and rates of convergence in normal approximation," *ArXiv*, vol. abs/1704.02658, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:13393096

[36] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman, "High-dimensional semiparametric Gaussian copula graphical models," *The Annals of Statistics*, vol. 40, no. 4, pp. 2293 – 2326, 2012. [Online]. Available: https://doi.org/10.1214/12-AOS1037

[37] L. Xue and H. Zou, "Regularized rank-based estimation of high-dimensional nonparanormal graphical models," *The Annals of Statistics*, vol. 40, no. 5, pp. 2541 – 2571, 2012. [Online]. Available: https://doi.org/10.1214/12-AOS1041

[38] P.-L. Loh and X. L. Tan, "High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination," *Electronic Journal of Statistics*, vol. 12, no. 1, pp. 1429 – 1467, 2018. [Online]. Available: https://doi.org/10.1214/18-EJS1427

[39] S. Mendelson and N. Zhivotovskiy, "Robust covariance estimation under $L_4 - L_2$ norm equivalence," *The Annals of Statistics*, vol. 48, no. 3, pp. 1648 – 1664, 2020. [Online]. Available: https://doi.org/10.1214/19-AOS1862

[40] E. Romanov, G. Kur, and B. Nadler, "Tyler's and maronna's m-estimators: Non-asymptotic concentration results," *Journal of Multivariate Analysis*, vol. 196, p. 105184, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0047259X23000301

[41] J. Lu, F. Han, and H. Liu, "Robust scatter matrix estimation for high dimensional distributions with heavy tail," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 5283–5304, 2021.

[42] K. Lounici, "High-dimensional covariance matrix estimation with missing observations," *Bernoulli*, vol. 20, no. 3, pp. 1029 – 1058, 2014. [Online]. Available: https://doi.org/10.3150/12-BEJ487

[43] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, 09 2018.

[44] A. Beck, *First-Order Methods in Optimization*. SIAM, 2017.

[45] T. T. Cai and H. H. Zhou, "Minimax estimation of large covariance matrices under ℓ_1-norm," *Stat. Sin.*, vol. 22, pp. 1319–1378, 2012.

[46] ——, "Optimal rates of convergence for sparse covariance matrix estimation," *Ann. Statist.*, vol. 40, no. 5, pp. 2389–2420, 2012.

[47] Y. Pitcan, "A note on concentration inequalities for u-statistics," 2019.

[48] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:268925835